Adult Income Prediction Project Proposal

By

Razan Jaad

Data Science Bootcamp

SDAIA Academy

[September - December 2021]

Abstract:

The target class in this project is adults. The dataset used in this project dataset contains 48842 rows and 14 attributes related to adults and the project makes some preprocessing for this dataset and uses logistic regression to analyze this data and make validation and testing.

The validation accuracy is 78.9%.
The testing accuracy is 79.2%

Society includes a wide range of individuals whose life tasks and priorities differ, but they have one thing in common, which is the interest in providing an adequate source of income to meet their needs. Many people may face difficulty in managing their money and their source of income may exceed a certain value, but with poor management, they do not know. Several methods based on machine learning have been proposed to deal with the problem of the process of predicting the value of income for an adult.

In this project, we present an empirical comparison of the most widely used machine learning approaches in revenue forecasting. Our experience involves using a logistic regression model.

We applied the above model in a standard data set that includes personal information about an adult, where this data is processed through prior data processing processes. The results show that the validation accuracy is 78.9%. and the testing accuracy is 79.2%.

Design:

Through Preprocessing Data and using Machine Learning techniques, this project has been found, which have a set of Continues and non-Continues for adults located all over the country. Accurate categorization of cases enables us to easily determine the income of adults.

Data:

The dataset contains 48842 rows and 14 attributes with a size equal to 5,202KB. The target column is Income. This dataset is used to predict and determine whether a person makes over 50K a year, dataset contains information for an adult to determine the income if <=50k $ or not. this data is processed by Preprocessing operations (https://archive.ics.uci.edu/ml/datasets/Adult)

Algorithm:

The algorithm we followed has two stages:

1. Processing the entered data, the input of this algorithm is our Dataset (70% training and 30% testing). This data is processed by Data Preprocessing, which works to convert raw data into ready-made data, which is the most important process that ensures the coordination of large data sets in a way through

which the existing data can be interpreted, one of the most important preprocessing operations is the data cleaning process (filling in missing values, deleting duplicate lines, etc..).

2. After processing, the data is entered into one of the models that we will know in the next paragraph, we are watching the output.

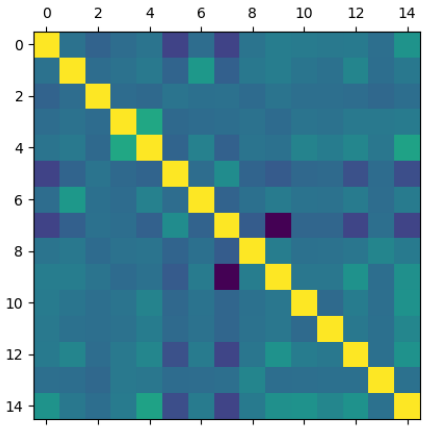
Models:

- logistic regression model: to classify the income of adults. The following image shows the building of the model.
    - DATA PREPROCESSING:
        1. Removing null values: Using the following code will remove all null values in the dataset.

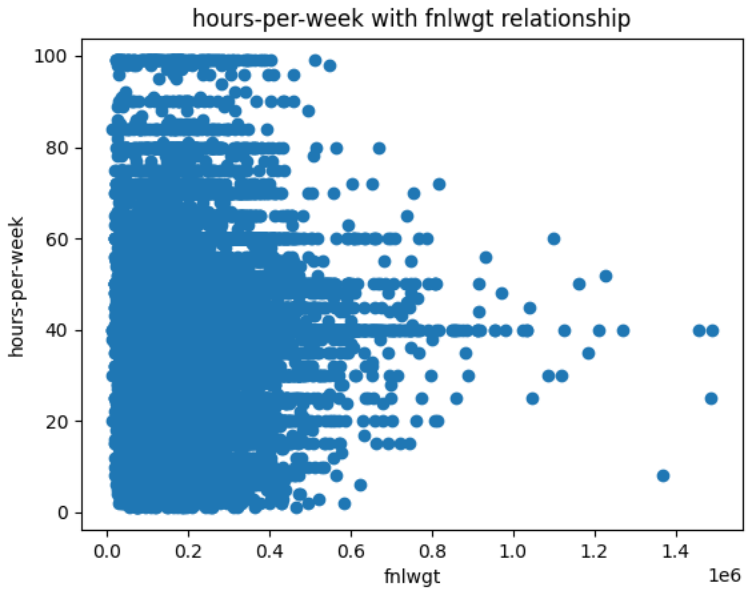
            Data = data.dropna()
        2. After that, All non-numerical types in the dataset were convert to numerical types by using the following code:

            ll = LabelEncoder()
            data["workclass"] = ll.fit_transform(data["workclass"])
            data["income"] = ll.fit_transform(data["income"])
            data["native-country"] = ll.fit_transform(data["native-country"])
            data["education"] = ll.fit_transform(data["education"])
            data["marital-status"] = ll.fit_transform(data["marital-status"])
            data["occupation"] = ll.fit_transform(data["occupation"])
            data["relationship"] = ll.fit_transform(data["relationship"])
            data["race"] = ll.fit_transform(data["race"])
            data["gender"] = ll.fit_transform(data["gender"])
    - DATA VISUALIZATION:
        1. First, the project makes all relationships between all columns using the code:
            plt.matshow(data.corr())
            plt.show()

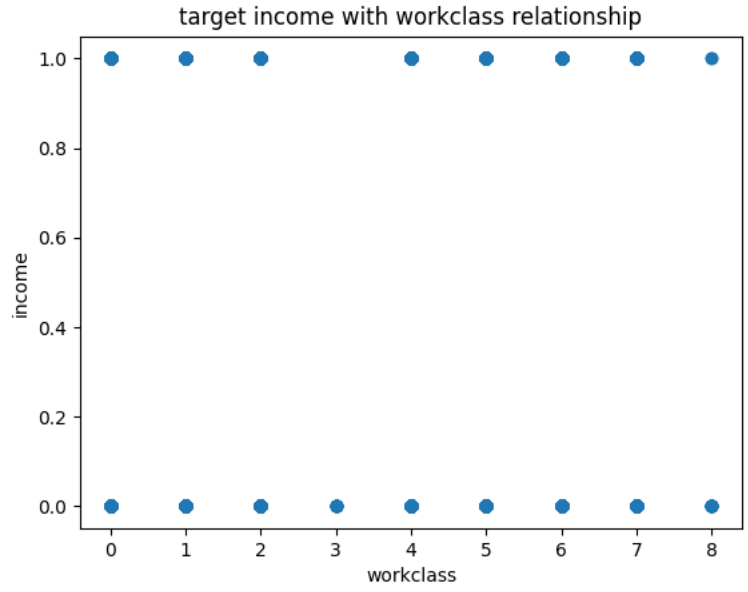
            - Each number in the image represents the index of columns in the dataset:



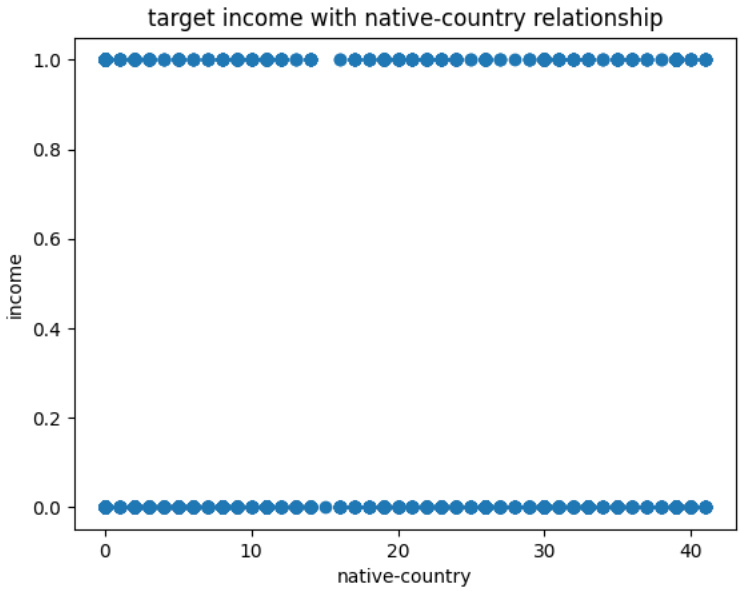
- Correlation between hours per week and fnlwgt:



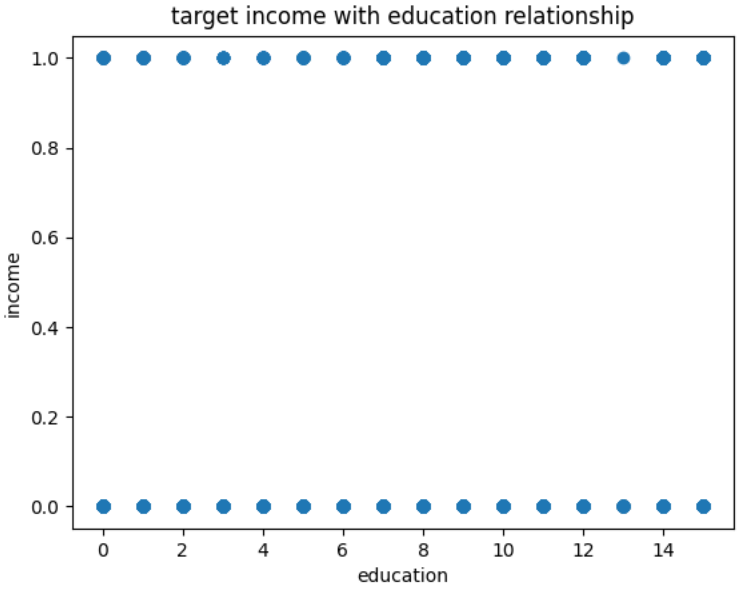
- Correlation between work class and income:



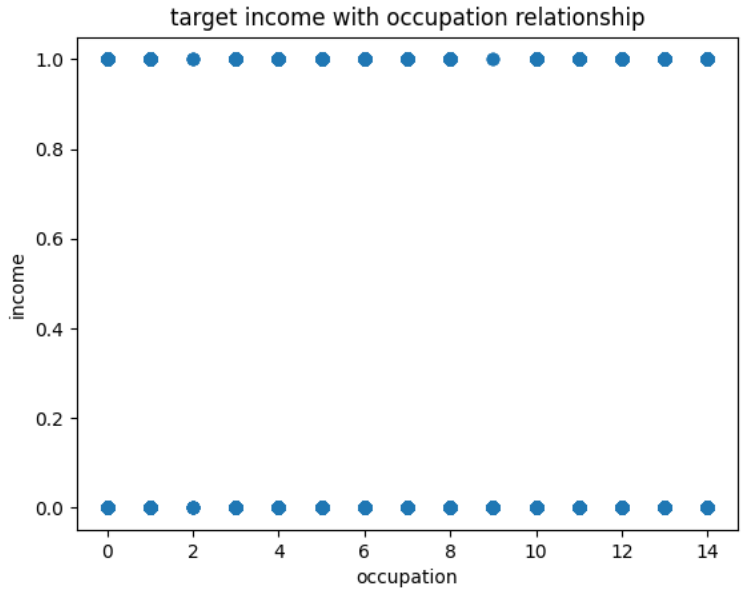
- Correlation between native country and income:



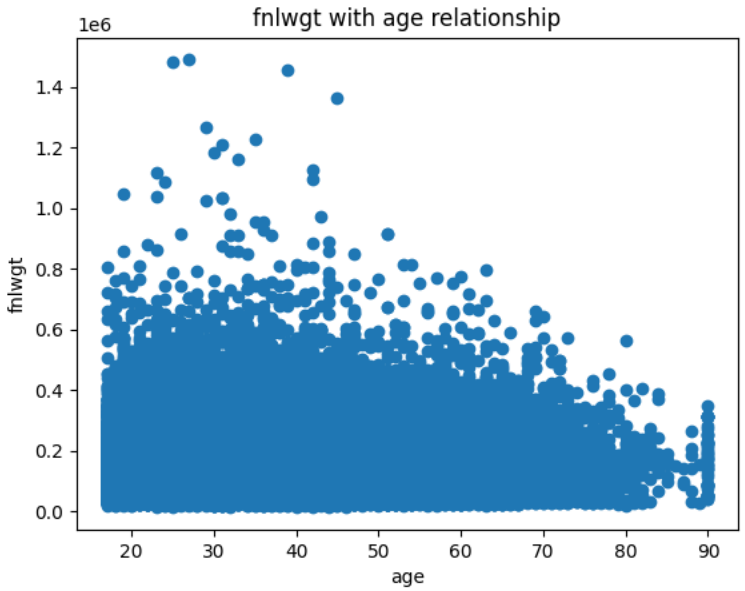
- Correlation between native education and income:



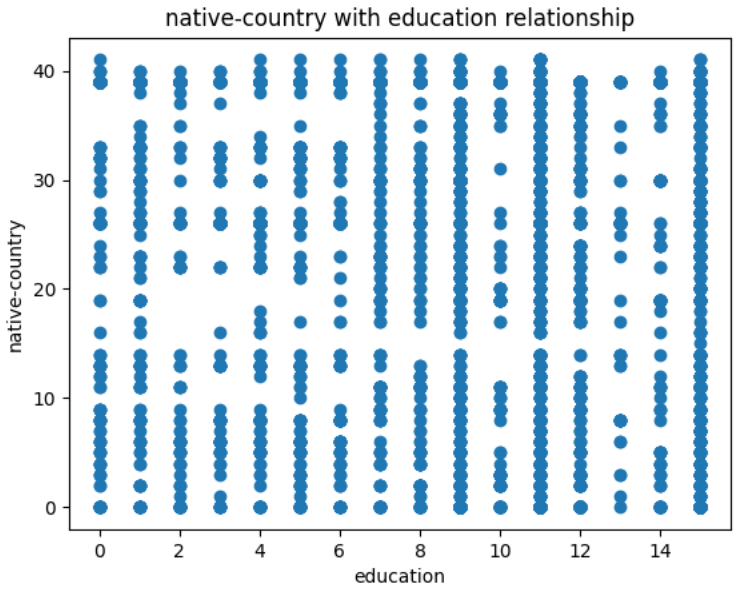
- Correlation between native occupation and income:



- Correlation between age and fnlwgt:



- Correlation between education and native country:



Tools:

1) Python: High-level Software language.
2) Jupyter notebook: Open-source web application.
3) Sklearn: Software library for modelling.
4) Pandas: Software library for data manipulation.
5) Matplotlib and Seaborn for plotting.

Conclusion:

The selected model gets about 79% accuracy after making some preprocessing steps and calculating the correlations between features. There are very small correlations between features, and this is the reason for the small accuracy.